Detection of Multiple Change Points from Clustering Individual Observations

JOE H. SULLIVAN

Mississippi State University, Starkville, MS 39762

In the preliminary analysis, also called Stage 1 analysis or retrospective analysis, of statistical process control, one may confront multiple shifts and/or outliers, especially with a large number of observations. This paper addresses the analysis of individual observations, and shows that the X-chart and CUSUM chart may fail to detect the presence of any shifts or outliers when multiple shifts and/or outliers are present. A new method is introduced which is effective in detecting single or multiple shifts and/or outliers. The algorithm and an effective stopping rule that controls the false detection rate are described. Suggestions are given for reducing masking and for diagnosing the number of shifts or outliers present.

Introduction

Much of the work in statistical process control (SPC) has focused on the "known parameters" problem, also called the prospective or "Phase II" problem, of rapidly detecting the presence of special causes of variation when the true in-control (IC) process parameters are accurately estimated or known. However, every process monitoring situation has an initial phase in which the IC parameters are unknown and must be estimated. Often, a Phase I, Stage 1 analysis, also called a retrospective or historical analysis, is used in which the process is operated for a while, and the resulting observations are collected and analyzed as a batch. The Stage 1 analysis involves simultaneously detecting the presence of special causes of variation and estimating the IC process parameters. Especially when it is appropriate to analyze individual observations, the presence of special causes can impair the parameter estimation, in turn masking their presence. Retrospective analysis is fundamentally different from prospective monitoring because of the masking problem and because there is no need to quickly signal an out-of-control (OC) condition. The retrospective stage of SPC closely resembles the change-point problem of statistical analysis.

Detecting one or more change points in a batch of observations has attracted considerable investigation in the statistical, engineering, and econometric literature. The general problem is as follows. Suppose there is an ordered sequence of observations (possibly multivariate), usually, but not necessarily, taken at equally spaced times. There is a change point between two successive observations if their statistical distributions are different. Between change points, the distributions are usually considered to be identical. Sometimes a model includes a deterministic change in the parameters over time, such as a linear trend, and a change point would be defined by an abrupt change in one or more parameters of the model rather than in the constantly changing statistical distribution itself. It is also common to refer to the last observation before the shift as the change point, although, strictly speaking, the change point comes at some time between the successive observations. Change points in the distribution of multivariate observations are also studied by Sullivan and Woodall (2000), and shifts in the regression parameters that define the relationship of multivariate observations are studied by Quandt (1958, 1960, and 1972). See Krishnaiah and Miao (1988), Zacks (1991), Barry and Hartigan (1993), and Lai (1995) for overviews.

Vostrikova (1981) and others have pointed out that a method for detecting a single change may be able to detect multiple changes by binary segmentation. If a change is detected, then the data are divided at the most likely location for a single change, and the change-point procedure is applied to each new group of data. This process is repeated until

Dr. Sullivan is an Associate Professor. He is a Member of ASQ. His e-mail address is jsullivan@cobilan.msstate.edu.

no group shows evidence of a change. For binary segmentation to be successful, it must be possible to detect the presence of multiple changes. However, when multiple change points are present, the series of observations need not follow any single model or regime, so that parameter estimation may be inaccurate. The presence of multiple changes may also impair the location estimator, so that the data are not segmented at the proper location.

McGee and Carleton (1970) suggest the multiple change point model

$$y_i = \mathbf{X}_i'\boldsymbol{\beta}_j + \varepsilon_i, \quad T_{j-1} < i \le T_j,$$
$$j = 1, ..., (R+1), \quad i = 1, \ldots, m,$$

where y_i is the i^{th} observed value, $\mathbf{X}_i$ is a p vector of predictor variables, $\boldsymbol{\beta}_j$ is a parameter vector, T_j is one of the R change points, $T_0 = 0$, and $T_{R+1} = m$, where m is the number of observations. The errors are normal and independent with zero mean and constant variance. McGee and Carleton (1970) describe their algorithm as a "wedding of hierarchical clustering and standard regression theory," and it operates as follows.

The algorithm begins at what is called level 1 by considering all possible clusters of consecutive observations having exactly the minimum size, $p+1$. The goodness of fit measure for a cluster defined by its first observaion, i_0, and number of observations, n_c, is

$$\phi = \frac{1}{n_c - p} \sum_{i=i_0}^{i_0+n_c-1} \hat{\varepsilon}_i^2,$$

The cluster minimizing ϕ is "fixed," meaning that these observations remain clustered in all subsequent steps. At the next step, level 2, all possible clusters of $p+1$ consecutive observations are considered, but eliminating from consideration those that include any observation that is part of an already fixed cluster. Two additional clusters are considered, the level-1-fixed cluster augmented by the adjacent previous observation and the level-1-fixed cluster augmented by the adjacent subsequent observation (unless the level-1-fixed cluster includes the first or last observation). Among those choices, the cluster minimizing ϕ is fixed, and the algorithm proceeds to the next step.

At each step all possible new clusters of minimal size are considered, subject to the constraints that the observations must be consecutive and must include no observation that is part of an already fixed cluster. In addition, all possible clusters formed by augmenting an already fixed cluster in either direction are considered. Sometimes two fixed clusters will be adjacent, so the minimum augmentation is more than one observation, since once observations are "fixed" into a cluster they cannot be separated in subsequent steps. The cluster minimizing ϕ is fixed at this step. The process continues in this way until all observations are joined into a single cluster or the stopping rule is satisfied. Satisfying the stopping rule signals the detection of one or more change points, the locations of which are estimated by the remaining boundary or boundaries.

The cluster fixed at each step will be one of four types: (a) a new cluster of minimal size; (b) an existing cluster extended to include the adjacent previous observation; (c) an existing cluster extended to include the adjacent subsequent observation; and (d) a merging of two adjacent fixed clusters. McGee and Carleton (1970) suggest stopping at the first type (d) combination with a significant F-test statistic, using the F-test for a single regression model vs. separate models. Thus, they regard that boundary and all other remaining boundaries as estimated change point locations. McGee and Carleton (1970) note that

> Such a decision rule might be modified according to how many isolated points remain. For example, we might decide ahead of time that we will not stop the hierarchical clustering until all but two of the original points are included in fixed clusters. As soon as all but two of the original points are thus included, we examine the probabilities associated with the F values for Type (d) clusters. The particular stopping rules adopted will reflect the best judgment of the investigator.

The hierarchical clustering algorithm can be viewed as a specialization of the EM algorithm proposed by Dempster, Laird, and Rubin (1977).

Hawkins (1976) generalizes the McGee and Carleton (1970) model to

$$y_t = f_j[t] + \varepsilon_t, \quad T_{j-1} < t \le T_j, \quad j = 1, ..., (R+1),$$

and reviews the literature in the context of this model. With normally distributed, zero mean errors, he distinguishes different situations depending on: (a) whether or not continuity is imposed on $f_j[T_j]$ and $f_{j+1}[T_j]$ for some or all change point locations T_j; (b) whether or not the errors have the same variance in all segments (homoscedasticity); and (c) whether the maximum number of change points is limited to one or not.

Hawkins (1976) briefly addresses the issue of estimating the number of change points present, following McGee and Carleton (1970) in using a sequence of significance tests on adjacent clusters. Hawkins notes the issue of multiple comparisons, writing that "the test statistics represent the most significant of a number of possible splits, and so should be interpreted conservatively." However, the estimation rule is not clearly stated.

Hawkins (1976) sets forth two alternative solutions, the first of which uses dynamic programming to find the exact change point locations that maximize the likelihood, subject to the restriction that all segments have at least $p + 1$ observations, one more than the number of parameters to be estimated in f_j. This restriction, required to make the likelihood finite, precludes the identification of clusters smaller than $p + 1$ and so impairs the effectiveness of detecting outliers. Hawkins gives an efficient algorithm for the exact solution that he considered, at that time, suitable for up to about 300 observations. Hawkins (2001) extends the dynamic programming algorithm to finding the exact maximum likelihood solution for any distribution of the exponential family. Although this algorithm is computationally fast enough to be applied to a large number of observations with only a personal computer, it still has the minimum size restriction that impairs outlier detection.

Even with an efficient dynamic programming algorithm, Hawkins (1976) notes that the computational requirements increase with the square of the number of observations. He proposes an alternative "hierarchical" solution whose computational requirements are proportional to the product of the number of observations and the number of change points. This method is computationally feasible for a much larger number of observations. Furthermore, clusters as small as a single observation are permitted, suggesting that the hierarchical approach may be better suited to the detection of outliers.

As Hawkins (1976) points out,

> . . . there is not necessarily any connection between the MLEs of the change points of an r - 1 segment model and those of an r segment one. However, in practice one would expect that if a change point were "real", its MLE would be stable as the number of segments fitted was increased. This reasoning suggests that one might impose the constraint that the r segment model utilizes the change points of the r - 1 segment model together with one fresh change point.
>
> With these additional requirements, it is clear that the solution will proceed hierarchically, and two methods of solution are by merging successively (as proposed by McGee and Carleton) and by splitting successively. The solution proposed here is a marriage of these techniques in which, at every iteration, each segment is examined to see whether it can be split into two significantly different segments, and each pair of adjacent segments is examined for the possibility of merging. This adaptation of the two strictly hierarchical methods yields a solution that is not necessarily hierarchical, and hence avoids the defect that a change point selected at an early stage of the analysis may lose its importance at a later stage when the segments on either side of it have been subdivided.

Although it is true that the hierarchical method can proceed either by merging, splitting, or a combination, there is a reason to prefer merging. Splitting corresponds to binary segmentation and is more susceptible to masking. If multiple change points are present, then the series, taken as a whole, may not follow the model for any set of parameters. Furthermore, it may not be possible to partition the series into two segments such that each segment follows the model. Thus, splitting and binary segmentation may fail to recognize the presence of any change points at all. On the other hand, the merging approach joins only those observations that conform well to a specific model. At each step, the model parameters are more accurately estimated for each cluster. The more accurate parameter estimation mitigates the masking problem and makes the identification of change point locations more accurate, and this logic is supported by simulation using a simple model with multiple step shifts in the mean. The defect of hierarchical clustering noted by Hawkins (1976) is well recognized when the clusters are not constrained by some ordering of the observations, in which situation a fundamental issue is estimating the parameters appropriate for each cluster. In that context, it is helpful to allow early clusters that were formed with poorly estimated parameters to be adjusted later when the parameter estimates are apt to be more accurate. These considerations also seem to apply in the current situation with splitting, but not merging. Hawkins' observation above suggests that if there are R "real" change points, then merging would remove the other, non-significant boundaries first. Thus with merging there seems to be little benefit in considering the splitting of an already fixed cluster, but it would be simple to do so.

The models of Hawkins (1976) and McGee and Carleton (1970) may be generalized to consider regime shifts between more general time series models. For example, the observations may be vectors rather than scalars, the conditional mean vector need not be independent of past observations, and the error vectors need not be independent. Thus, a proposed, more general, change point model is

$$\mathbf{y}_t = \mathbf{g}_j[\mathbf{x}_t, \mathbf{w}_t], \quad T_{j-1} < t \le T_j, \quad j = 1, ..., (R+1),$$

where $\mathbf{y}_t$ is the p_n vector of observed variables for observation t, $\mathbf{x}_t$ is a p_p regressor vector for observation t, $\mathbf{w}_t$ is a p_s vector of possibly unobserved state variables at observation t, and $\mathbf{g}_j$ gives the model for regime j. The vector autoregessive model (Litterman (1986)) is obtained with

$$\mathbf{g}_j[\mathbf{x}_t, \mathbf{y}_t] = \mathbf{f}_j[\mathbf{x}_t] + \Phi_j[L]\ \mathbf{y}_t + \varepsilon_t,$$

where $\Phi_j\ [L]$ is the lag operator of order p_L, which includes $p_L - 1$ distinct $p_n \times p_n$ parameter matrices, and ε_t is a p_n random vector with expectation zero. The model may be simplified to a linear model with $f_j[\mathbf{x}_t] = \mathbf{B}_j\mathbf{x}_t$, where $\mathbf{B}_j$ is a p_n by p_p parameter matrix. The errors may be uncorrelated but possibly heteroscedastic with uncorrelated errors specified by

$$E[\boldsymbol{\varepsilon}_t\ \boldsymbol{\varepsilon}'_\tau] = \begin{cases} \Omega_j & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

Alternatively, the more general conditional heteroscedastic model may be used, in which case the conditional covariance matrix is specified as an ARMA$[p_a, p_m]$ model of the innovation matrices $\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}'_t$ with the model parameters constant between change points.

The model can be written as a vector structural equations model (Giannini, 1992)

$$\begin{aligned} \mathbf{g}_j[\mathbf{x}_t, \mathbf{w}_t] &= \boldsymbol{\Gamma}_j\mathbf{w}_t \\ \Phi_j[L]\mathbf{w}_t &= \mathbf{f}_j[\mathbf{x}_t] + \varepsilon_t, \end{aligned}$$

where $\boldsymbol{\Gamma}_j$ is a $p_n \times p_s$ observation matrix. The structural model can include the conditional heteroscedastic error model.

Instead of detecting multiple changes one at a time, an alternative is to detect them all at once. The number of change points can be viewed as a model parameter and estimated by a penalized likelihood method that maximizes $L_q - Fk_q$, where q indexes the alternative models, L_q is the maximized likelihood for model q, k_q is the number of parameters in model q, and F is the dimensionality penalty. Two of the most commonly used criteria are the Akaike information criterion (AIC, Akaike (1974)), corresponding to $F = 1$, and the Schwarz information criterion (SIC, Schwarz (1978)), also called the Bayesian information criterion (BIC), corresponding to $F = 0.5\ell\text{n}[m]$, where $\ell\text{n}[\cdot]$ is the natural logarithm. In comparing these approaches, Diebold (2001, p. 87) states that when the true model is fixed, the SIC is consistent but the AIC is not, while the AIC is asymptotically efficient but the SIC is not if the true model dimensionality increases with m in a specified way. The SIC penalizes additional parameters more heavily, except for small m, and so tends to select a more parsimonious model. Many other criteria have been proposed, including adaptive choices of F. See George (2001) for further discussion.

Yao (1988) established some properties of the SIC in detecting multiple change points and in estimating their number and locations. However, his analysis assumed knowledge of the change point locations that maximized the likelihood function for each possible number of changes. As Hawkins (1976) shows, finding the exact maximum likelihood locations becomes computationally infeasible with a large number of observations. Sullivan (2002) discusses the SIC in this context, using hierarchical clustering to estimate the maximum likelihood locations and showing the weakness of binary segmentation in detecting the presence of multiple change points. For example, the Chernoff and Zacks (1964) test, although effective in detecting a single change point, is shown to be ineffective in detecting the presence of two step shifts in the mean. The lack of sensitivity extends to many other configurations of multiple change points. A comparison of the results in Sullivan (2002) with those of this paper shows that the method proposed here is more accurate in controlling the false detection probability and in estimating the number of change points. Therefore, the SIC is not considered further in this article.

The array of possible change point models is overwhelming. It seems reasonable to make a beginning by considering the simplest possible multiple change point situation, univariate observations with independent, identically distributed (i.i.d.) normal errors and a constant mean between change points, which is described in the next section.

Then, the following section gives a numerical example with four shifts to demonstrate the proposed clustering algorithm and its advantage compared with an X-chart (Shewhart chart) and a CUSUM chart. Next, the sequence of distances is viewed as a single multivariate observation, and its expected value is estimated by simulation for the IC and several alternative OC situations. The proposed clustering algorithm is shown to have a more uniform detection probability than the CUSUM and X-charts. The CUSUM is not effective in detecting outliers in any number, and the X-chart is not effective in detecting a small number of shifts.

Methodology with a Simple Model

Suppose there are m independent observations, $x_1, x_2, \ldots, x_m$, from one or more univariate normal distributions all having the same variance σ^2. There are R shifts in the mean, and the shift locations are T_r, $r = 1, \ldots, R$, subject to $0 < T_1 < \ldots < T_R < m$. Let θ_i represent the mean of observation x_i, and define $T_0 = 0$ and $T_{R+1} = m$. Then, $\theta_i = \mu_r$, for $T_{r-1} < i \leq T_r$, for $r = 1, \ldots, R+1$, subject to $\mu_r \neq \mu_{r+1}$. We want to determine if the process is IC, which corresponds to $R = 0$. If the process is not IC, we can obtain diagnostic information by estimating the number of shifts and their location(s).

The clustering algorithm starts with $m-1$ boundaries separating the observations into singleton clusters. The boundaries are indexed by k_j, and associated with each boundary is a location l_k, the last observation in the cluster, and a distance d_k, measuring the dissimilarity of the means of its adjacent clusters. The absolute value of the Student's t-statistic for a difference in two means can be used, which is given by

$$d_k = \frac{|\bar{x}_k - \bar{x}_{k+1}|}{s\sqrt{\frac{m_k+m_{k+1}}{m_k m_{k+1}}}} \tag{1}$$

where m_k and m_{k+1} are the number of observations in the adjacent clusters, $\bar{x}_k$ and $\bar{x}_{k+1}$ are the sample means, and s is an estimate of the common standard deviation of all clusters. The ranking of the distances does not depend on the value of s, so without any loss of generality the value $s = 1$ is used.

At the start of step K, $K = 1, \ldots, m-1$, there are $m - K$ boundaries. The one with the smallest distance ($k^* = \arg\min\,[d_k]$) is removed, and the remaining distances are updated. The location of the removed boundary ($l^*_{m-K} = l_{k^*}$) and its distance ($d^*_{m-K} = d_{k^*}$) are saved. The process continues until all the boundaries are removed. Note that the sequence $\{d^*_i\}$ begins with the distance of the last-removed boundary, d^*_1.

Logically, if the process is IC, then the sequence $\{d^*_i\}$ should decrease slowly and smoothly. If there are, say, R large shifts, then $d^*_R - d^*_{R+1}$ should be "large," and the distances should decrease slowly beyond d^*_{R+1}. These distances can be used in a decision rule for recognizing the presence of multiple shifts, as discussed later.

Sullivan (2002) notes that the hierarchical clustering with the simple model of iid normal errors and step shifts in mean has an equivalent regression model. Let $\mathbf{x} = \mathbf{U}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$, where $\mathbf{x}$ is the vector of m observations, $\boldsymbol{\varepsilon}$ is a vector of m independent, standard normal errors, and $\mathbf{U} = [u_{i,j} = I_i[j]] = [\mathbf{u}_{(1)} \cdots \mathbf{u}_{(m)}]$ is an $m \times m$ matrix. The indicator function $I_i[j] = 1$ if $j \leq i$ and 0 otherwise. The vector $\boldsymbol{\beta}$ gives the shifts in the mean, $\beta_{i+1} = \theta_{i+1} - \theta_i$ for $i \in \{T_1, T_2, \ldots, T_R\}$, and 0 otherwise. Backward elimination of the variables $\mathbf{u}_{(i)}$ from this model gives the same sequence as the merging direction for the simple model. Forward selection, which corresponds to splitting, was found (using simulation) to be less accurate, due to masking.

One of the main difficulties in the Stage 1 analysis with individual observations is accurately estimating σ in the presence of multiple shifts and/or outliers. Typically, the average of the moving ranges is used. It is of interest to construct an alternative robust estimator and compare it with the estimator based on the average of the moving ranges and the sample standard deviation. Suppose there are n_s shifts and n_o outliers. Then, if the hierarchical clustering works as intended, the clusters would consist of homogeneous observations with standard deviation σ until fewer than $n_s + 2n_o$ boundaries remain. This suggests that some suitable fraction of the observations, such as $0.8m$, could be chosen in advance, and, when that number of boundaries had been removed, the sum of squares within clusters could be used to form a robust estimator of σ. The optimal fraction to use depends on the dimension, the number of observations, the maximum number of shifts and outliers that may be present, and the objective criterion. To initially explore this topic, a robust estimator was defined and calculated as a part of the simulations at the point where $0.2m$ boundaries remain.

The robust estimator used was

$$s_r^2 = \frac{1}{m-K-1} \sum_{k=1}^{K+1} \sum_{i=1+T_{k-1}}^{T_k} (x_i - \bar{x}_k)^2,$$

where $K = 0.2m$, rounded to the nearest integer, and the mean of cluster k is

$$\bar{x}_k = \frac{1}{T_k - T_{k-1}} \sum_{i=1+T_{k-1}}^{T_k} x_i, \quad 1 \leq k \leq K+1.$$

Simulation is used to estimate the ratio of the expected value of s_r^2 to the true variance with IC data, and this ratio is used to remove the bias. Later, the performance of this estimator is compared with the estimator based on the average of the moving ranges and the sample standard deviation.

TABLE 1. Observations and Calculated Statistics for the Example

i	X_i	CUSUM$_i$	l_i^*	d_i^*	i	X_i	CUSUM$_i$	l_i^*	d_i^*
1	1.190	1.117	48	3.341	31	1.630	2.503	30	1.142
2	1.230	2.277	37	5.382	32	1.590	4.044	52	0.805
3	1.510	3.734	23	4.237	33	0.407	4.332	51	1.118
4	−0.531	3.027	15	5.876	34	−0.128	4.053	56	1.071
5	−0.705	2.137	10	2.622	35	1.840	5.859	17	0.988
6	0.429	2.448	6	1.934	36	0.346	6.082	12	0.592
7	1.600	4.000	3	2.705	37	−0.104	5.828	13	0.827
8	1.830	5.795	54	1.702	38	−1.850	3.725	58	0.640
9	1.560	7.305	53	2.309	39	−0.410	3.147	43	0.607
10	0.661	7.862	32	1.756	40	−0.750	2.209	55	0.552
11	−0.727	6.948	29	2.287	41	−1.950	−0.001	33	0.529
12	−0.536	6.237	35	0.897	42	0.330	0.205	36	0.445
13	0.210	6.316	34	1.942	43	−1.460	−1.485	2	0.343
14	−0.626	5.509	41	0.711	44	−0.757	−2.430	44	0.339
15	0.933	6.354	42	1.891	45	−1.100	−3.739	39	0.336
16	−1.580	4.537	21	1.551	46	−2.210	−6.224	28	0.289
17	−1.490	2.815	20	1.195	47	−0.502	−6.899	26	0.268
18	−2.400	0.128	18	1.721	48	−0.752	−7.840	47	0.247
19	−0.710	−0.767	14	1.692	49	1.380	−6.521	8	0.177
20	−0.691	−1.643	22	1.652	50	1.340	−5.245	7	0.227
21	−2.310	−4.234	27	1.650	51	0.100	−5.282	11	0.189
22	0.210	−4.155	24	1.482	52	1.230	−4.122	4	0.172
23	−1.460	−5.845	40	1.147	53	−0.040	−4.308	25	0.148
24	0.210	−5.766	38	1.450	54	2.610	−1.686	57	0.102
25	1.280	−4.553	50	1.425	55	0.379	−1.428	16	0.089
26	1.430	−3.182	46	1.219	56	0.937	−0.579	1	0.040
27	1.590	−1.640	45	1.338	57	−0.176	−0.909	49	0.040
28	0.252	−1.517	59	1.240	58	−0.279	−1.348	31	0.040
29	−0.040	−1.703	9	1.214	59	0.333	−1.139	19	0.019
30	2.610	0.919	5	1.196	60	1.210	0.000		

Numerical Example

An example with four shifts is analyzed to illustrate the clustering algorithm. This example was chosen because it presents visually compelling evidence of multiple shifts, which are detected by the proposed clustering method. However, the X-chart and CUSUM charts do not signal, despite the strong visual evidence of four shifts. The 60 observations to be analyzed are listed in the second column of Table 1, and the X-chart is shown in Figure 1. For this data the centerline is the sample mean = 0.135, and the estimated standard deviation based on the average moving range is $\overline{\text{MR}}\ /d_2 = 1.39/1.13 = 1.23$. Control limits are shown 3 standard deviations from the mean, as well as limits for zones that are 1 and 2 standard deviations from the mean. The largest deviation from the centerline is at observation 18 and is well within the control limits. Thus, the X-chart (without any supplementary runs rules) gives no signal.

Another widely used control chart is the CUSUM chart, which is shown in Figure 2 and also fails to signal. The plotted values for the CUSUM chart are given in column 3 of Table 1. As recommended by Sullivan and Woodall (1998), in this retrospective analysis the deviations of each observation from the sample mean are cumulatively summed, without subtracting a constant. In this example, the cumulative sums are all divided by the standard deviation, estimated from the moving ranges. The control limits are chosen by simulation to be ±8.59, which gives the same false alarm probability as the X-chart.

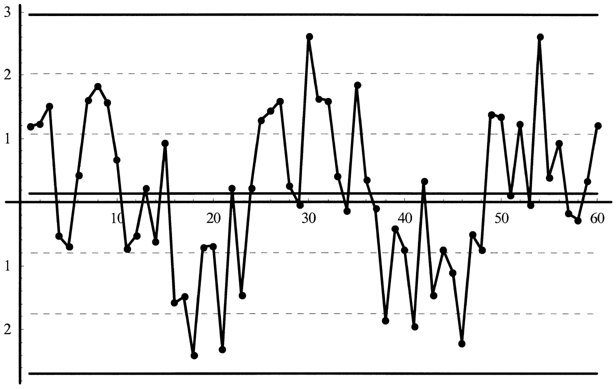

FIGURE 1. X-Chart for the Example.

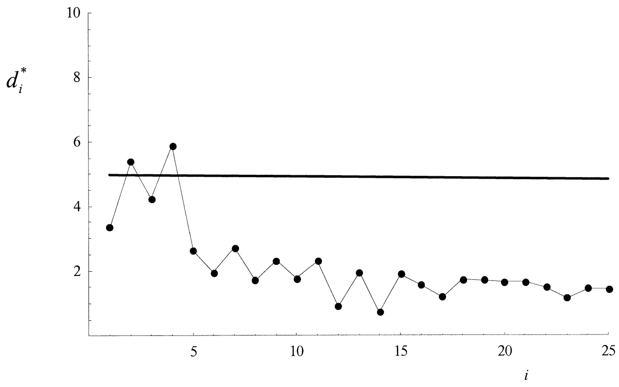

FIGURE 3. Distances $\{d_i^*\}$ for the Example.

Next, the boundary removal sequence of the clustering algorithm is calculated. Initially, 59 boundaries define 60 clusters of one observation each. The distance is calculated for each boundary according to Equation (1). The smallest distance is 0.019 which is associated with boundary 19, so $l_{59}^* = 19$ and $d_{59}^* = 0.019$. Boundary 19 is removed, merging observations 19 and 20, the distances are recalculated, and the algorithm proceeds to the next step. The sequences $\{l_i^*\}$ and $\{d_i^*\}$ are given in Table 1. A plot of $\{d_i^*\}$ is shown in Figure 3. As discussed later, to control the false detection probability in light of the multiple comparisons, only d_1^* and d_2^* enter into the proposed decision rule. One decision rule is based on the maximum of d_1^* and d_2^*, and a suitable maximum under this rule is shown in the figure. Since d_1^* and d_2^* are not both below the maximum, there is an indication of an OC condition. A more effective decision rule is given later.

When a signal is given, the chart can be further analyzed for diagnostic information. In this example, there is a sharp change between distances four and five, a pattern corresponding to four shifts in the mean. Thus, the plot can be interpreted to indicate the presence of four shifts, with estimated locations $\{l_i^*,\ i = 1, \ldots, 4\} = \{48, 37, 23, 15\}$. If the observations are divided at these locations, the p-values for the test for a difference in the means of adjacent clusters are, respectively, $\{0.0000152,\ 0.00000297,\ 0.00000687,\ 0.0000940\}$, confirming that these locations are reasonable in separating clusters with dissimilar means.

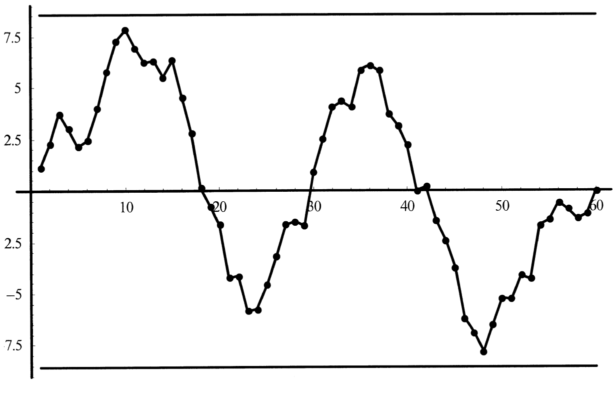

FIGURE 2. CUSUM Chart for the Example.

The stopping rule proposed by McGee and Carleton (1970) and Hawkins (1976) depends on the least significant (maximum) p-value at each of the steps, which is plotted in Figure 4. A large p-value indicates that one of the boundaries is not really a change point or valid change points are being masked. Thus, the largest K for which all p-values are significant can be taken as an estimate of the number of change points, and the data are OC if $K > 0$. As Hawkins (1976) notes, the significance should be interpreted conservatively because of the multiple comparisons. With a conservative significance level, such as 0.001, the largest K for which all boundaries are significant is 4. With a less conservative significance level, such as 0.02, then seven shifts would be indicated. However, the seven-shift solution is a finer partition of the four-shift solution, so that if the four shifts are real then their locations would be preserved in any solution with a larger number of shifts.

The large p-value for a single shift illustrates the masking problem—there may be no single location that divides the data into clusters having significantly different means, although division into a greater number of clusters having significantly different means is possible.

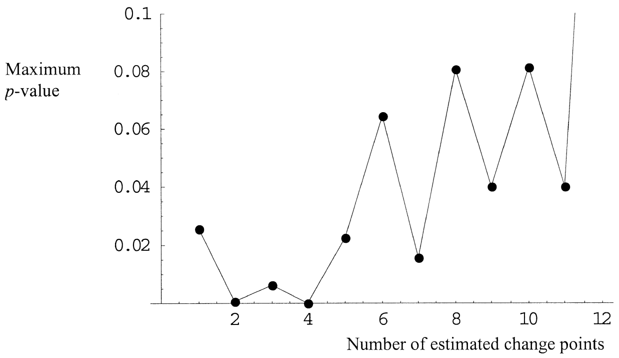

FIGURE 4. Maximum p-values.

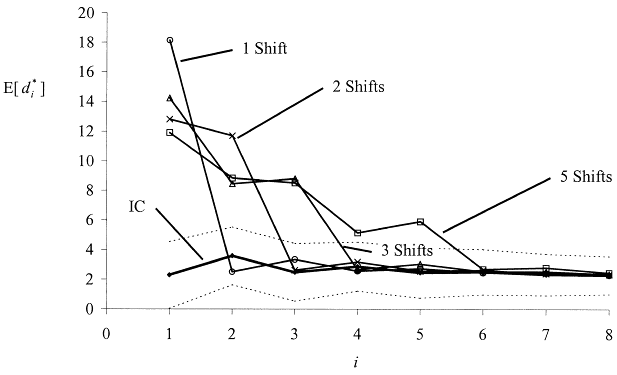

FIGURE 5. Expected Values of d_i^* with Shifts.

Hawkins (2001) discusses the inferential difficulties in estimating the number of change points. He points out that the null hypothesis of no shifts can be tested against the alternative of exactly k segments by a generalized likelihood ratio (GLR) test statistic, noting that the test statistic does not follow the expected asymptotic chi-squared distribution. Hawkins states that in the simplest case of normal data with constant variance and at most a single change point in the mean, there is not an asymptotic distribution for the GLR test statistic, which increases without bound with the sample size. He concludes that "the failure of conventional asymptotics in even this easiest case is an indication of the technical difficulty of the more general situation." Thus, there is a benefit in considering stopping rules that accurately control the false detection probability, the next topic.

Alternative Detection Rules

The inference from the sequence of boundary distances about the presence of special causes is enriched by viewing the sequence as a vector observation whose distribution depends on the data. It is useful to examine the effect of various OC situations on this multivariate distribution. We can start by considering the IC expected values of the first few values of $\{d_i^*\}$, estimated from simulation, as shown in Figure 5. The IC expected values are shown with the heavy solid line, with thin dashed lines two standard deviations above and below. For example, the IC expected values are 2.28 for d_1^* and 3.57 for d_2^*. It is somewhat surprising that the distance of last boundary to be removed, supposedly the most likely location of a single shift, does not have the largest expected value when the data follow the null distribution. Instead, it is the next-to-last-removed boundary whose IC average distance is largest. In addition to the IC expected values, there are lines for OC conditions with one, two, three and five equally spaced shifts. The shifts are all the same size, 2σ.

The curve with a circle symbol shows the result for one shift, which affects d_1^* the most. The average of d_2^* is decreased slightly, with minor perturbations in the others. With two shifts, shown with a cross symbol, d_1^* and d_2^* are increased, while the others are affected very little. Three shifts (triangle) affect the first three distances, and five shifts (square) affect the first five distances. These patterns are useful in diagnosis after an OC signal is generated.

Another OC model is the presence of one or more outliers. Figure 6 shows the OC expected values with one, two, or three equally spaced outliers. All of the outliers have the same mean, which differs from that of the adjacent observations by 12σ. The dashed line with a circle symbol corresponds to a single outlier, showing an increase in the mean of d_2^*, a much smaller increase in the mean of d_1^*, and hardly any change in the others. The plot for two outliers is solid with a cross symbol, showing that the means

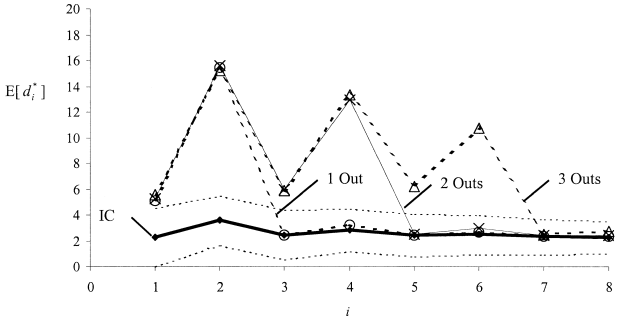

FIGURE 6. Expected Values of d_i^* with Outliers.

of d_2^* and d_4^* are increased. Three outliers are shown with the dashed curve with the triangle plot symbol, showing that the means of distances 2, 4, and 6 are elevated. The effect on d_1^* and d_2^* is nearly the same with 1, 2, or 3 outliers, so the plots nearly coincide in the first two places.

This plot suggests that pure outliers mostly affect the even-numbered distances. The rationale is that n_0 distinct outliers can be regarded as $2n_0$ shifts in the mean. Distances beyond $d_{2n_0}^*$ are not affected much, since they do not correspond to real change points. If there are no mean shifts then the clusters with $m - 2n_0$ boundaries remaining will all have about the same mean but will not have been merged yet because there is a single outlier between each larger cluster. Thus, all the boundary distances are large. After an outlier is merged with an adjacent cluster, the cluster mean changes, but only slightly. Thus, the distance associated with either of its boundaries is much smaller than the minimum distance at the previous step. This accounts for the sawtooth shape of the plot of $\{d_i^*\}$.

There is also an issue of effect size when comparing multiple shifts. The effect size, measured by Equation (1) with population parameters, depends on the size of the shift and the number of observations in each cluster. In Figure 5, with a constant number of observations and constant shift size, the effect size decreases with the number of shifts.

In detecting special causes, we next consider how to best use the single multivariate observation vector in a decision rule. Since each OC distribution creates a unique distribution of the distances, knowledge of a specific OC pattern, such as three shifts or five outliers, would sensibly guide the analysis process. Generally, there is no specific OC pattern to be detected, but rather the objective is the detection of any reasonable number of shifts and/or outliers. To create a decision rule, a summary statistic is needed. An important issue is how many boundary distances to use in the summary statistic. Since outliers in any number have little effect on the first distance, using it alone would not give much sensitivity in detecting outliers. On the other hand looking at all of them for the last significant d_i^*, as suggested by McGee and Carleton (1970) and Hawkins (1976), introduces the multiple comparisons problem and makes controlling the overall false detection probability difficult. Including the first two boundary distances in the summary statistic seems reasonable because at least one is affected by any number of shifts and/or outliers. Furthermore, the null distribution should change only slowly with the number of observations, simplifying the task of finding an approximate expression for the critical value. For these reasons, a summary statistic of the first two boundary distances was sought. However, there is a trade-off, since including a greater number can improve performance with many shifts/outliers at the expense of performance with a single shift/outlier.

We now focus on the bivariate vector $\{d_1^*, d_2^*\}$ and depict the effect of different OC situations on its statistical distribution. The IC and seven different OC situations are considered. The statistical distributions are depicted in Figure 7 by plotting 1000 samples from each distribution. These 8000 observations can be grouped visually into eight clusters, corresponding to the eight situations that characterize the primary data. The IC cluster is closest to the origin and shown with the plus symbol. The situations corresponding to one, two, or three equally spaced shifts and one centrally located outlier are labeled. The shift conditions all include shifts of size 2, and include size 1 shifts for one and two shifts. The outlier has shifts of size 8 and 16. The size is measured by Equation (1) with population parameters. The cluster corresponding to one shift of size 1 is centered at (11.57, 2.57), and it shows a large change in the first distance compared with the IC cluster, but little change in the second distance.

The OC distributions cluster farther from the origin than the IC distribution and in a direction that depends on the specific nature of the OC situation. For a specific configuration of outliers or shifts, the direction is roughly constant, and the mean vector is shifted out in proportion to the effect size. One shift moves the mean vector along a nearly horizontal line since only d_1^* is affected. Two shifts move the mean vector along a 42° line, with the cluster for two shifts of size 1 centered at (6.60, 8.94). Three shifts move the mean vector along the line $\mathrm{Tan}^{-1}[8.5/14.2] = 31°$. Since any reasonable number of equally spaced outliers has about the same effect on $\{d_1^*, d_2^*\}$, the displacement direction will not change much with the number of outliers and corresponds to the line $\mathrm{Tan}^{-1}[15.5/5.1] = 72°$.

In distinguishing between the IC and the various OC distributions, it may be tempting to consider all OC conditions as contributing to a single distribution, but that can be misleading. The correlation is negative for the IC distribution and also negative if all the OC observations are taken as a single dis-

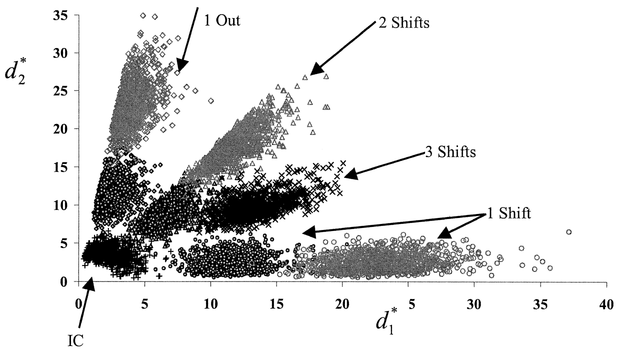

FIGURE 7. Samples of $\{d_1^*, d_2^*\}$ from Various Distributions.

tribution. However, for any specific OC distribution the correlation is positive.

So, how should we partition the bivariate space to classify the data as either IC or OC? Ten plausible alternative summary statistics were considered, but for brevity only two are described here. The one with the most uniform performance across all types of OC conditions studied was based on the statistic $\mathbf{c}_*' \boldsymbol{\Sigma}^{-1} \mathbf{c}_*$, where $\mathbf{c}_*$ is the vector $(\mathbf{c} - \boldsymbol{\mu})$ with negative elements replaced by zero, and $\mathbf{c} = (d_1^*, d_2^*)'$. The expected value and covariance matrix of $\mathbf{c}$ are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which are estimated from simulation. The decision rule based on this statistic was used in all subsequent performance comparisons. Another statistic, $\text{Max}[d_1^*, d_2^*]$, was used to determine the critical value for Figure 3.

Performance Comparison

The statistical distribution of the data can depart from the IC distribution in infinitely many ways. The model used here for multiple mean shifts is uniformly spaced shifts alternating between two means. For example, a single shift would be midway in the data, and two shifts would be located after one-third and two-thirds of the observations. An alternative model would be random shift locations, but then the shifts can resemble outliers if a shift is near the end of the data or another shift. The model for one or more outliers is equal spacing and equal mean shifts. The change in the mean was adjusted with the number of shifts or outliers to maintain the same effect size as defined by Equation (1) evaluated using population parameters. With R equally spaced shifts of equal magnitude in m observations, there are $m/(R+1)$ observations on either side of each shift, so the effect size for each shift is

$$\delta_M = \frac{|\mu_1 - \mu_2|}{\sigma} \sqrt{\frac{m}{2(R+1)}},$$

where μ_1 and μ_2 are the two means. For R equally spaced outliers, there are 1 and $(m-R)/(R+1)$ observations adjoining each shift, so the effect size is

$$\delta_O = \frac{|\mu_1 - \mu_2|}{\sigma} \sqrt{\frac{m-R}{m+1}},$$

The performance was compared with 64 observations simulated from normal distributions. The effect size was $\delta_M = 1.15$ for shifts and $\delta_O = 3$ for outliers. The values for $|\mu_1 - \mu_2|/\sigma$ are shown in Table 2. Without any loss of generality, $\sigma = 1$ and $\mu_1 = -\mu_2$ were used in the simulation.

Traditionally, performance comparisons of control charts for the prospective Phase II and Stage 2 of Phase I are conducted using the average run length (ARL). In the retrospective analysis, Stage 1 of Phase I, the ARL is not meaningful because an OC signal is generated for the entire batch of data. For retrospective performance comparison, it is reasonable to use the probability that a chart generates an OC signal, referred to as the signal probability. For a fair comparison, all charts should be adjusted

TABLE 2. Shift Sizes Used to Evaluate Signal Probabilities

Condition	$\lvert\mu_1 - \mu_2\rvert/\sigma$
IC	0
s01	0.288
s02	0.352
s03	0.407
s04	0.455
s05	0.498
s06	0.538
s07	0.575
s08	0.610
s09	0.643
s10	0.674
o1	3.024
o2	3.048
o3	3.073
o4	3.098
o5	3.124
o6	3.151
o7	3.154

TABLE 3. UCL Values for Control Charts

Chart	UCL
cusumSSD	8.38179
cusumMR	8.59497
cusumTV	8.41948
shewhartSSD	2.92160
shewhartMR	2.99504
shewhartTV	2.97686
$\mathrm{Max}[d_1^*, d_2^*]$	4.98551
$\mathbf{c}_*'\boldsymbol{\Sigma}\mathbf{c}_*$	4.10918

to have the same IC signal probability, referred to as the false detection probability. In selecting the common false detection probability, it is reasonable to use the false detection probability of a traditional X-chart with known parameters and three sigma limits, which is $1-(1-2*\Phi[-3])^m$, where Φ denotes the cumulative distribution function of the standard normal distribution. With 64 observations, the false detection probability is 0.159, which is the value used in the simulations. All signal probabilities were estimated with 100,000 simulations.

The performance comparisons include three versions each of the CUSUM chart and the X-chart. The charts differed in the estimation of the standard deviation, using the sample standard deviation (SSD), the average of the moving ranges divided by d_2 (MR), or the true value (TV). Of course, the true value of the standard deviation would not be known in practice, so it is not a practical analysis tool, but provides useful insight into the performance of the chart with the best possible estimator. For each version, the control limits were determined by simulation and are given in Table 3. The control limits for the X-chart with the true σ would be exactly 3, instead of 2.98, except for the slight inaccuracy from simulation. The UCL for $\mathrm{Max}[d_1^*, d_2^*]$ was used in Figure 3.

Figure 8 compares the signal probabilities of the clustering algorithm with those of the CUSUM charts using the three alternative estimators. Various data situations appear on the horizontal axis. The leftmost item is the IC situation. The next 10 correspond to 1 to 10 shifts, respectively, and the last 6 correspond to 1 to 6 outliers, respectively. The performance of the clustering algorithm is shown with the heavy solid line and a square symbol. For detecting a single shift midway in the observations, the CUSUM chart is unexcelled. However, as the number of shifts increases, the signal probability falls until, at 10 shifts, it is only a little more than the false detection probability. Furthermore, both of the feasible CUSUM charts are biased, in the sense that the signal probability with any number of outliers is less than the false detection probability. The advantage of the clustering algorithm is that it maintains a nearly uniform signal probability over all of the OC conditions.

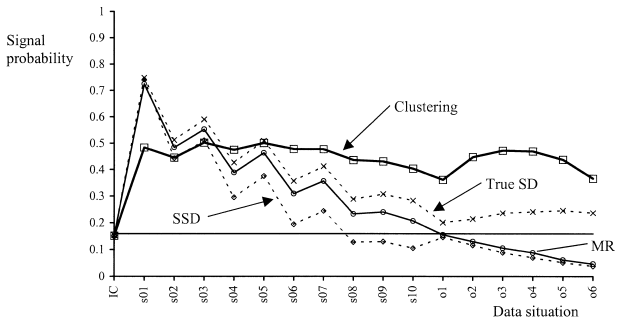

FIGURE 8. Signal Probability for CUSUM Chart and Clustering Algorithm.

The clustering algorithm performance is compared with that of the X-chart (without runs rules) in Figure 9. The X-chart does well in detecting outliers, but not very well in detecting a single shift. Again, the clustering method gives a more uniform performance over the collection of OC situations.

Accurate estimation of the variance is important in the performance of any of the charts. As mentioned earlier, it is of interest to see how s_r performs in estimating σ. It is compared with the sample stan-

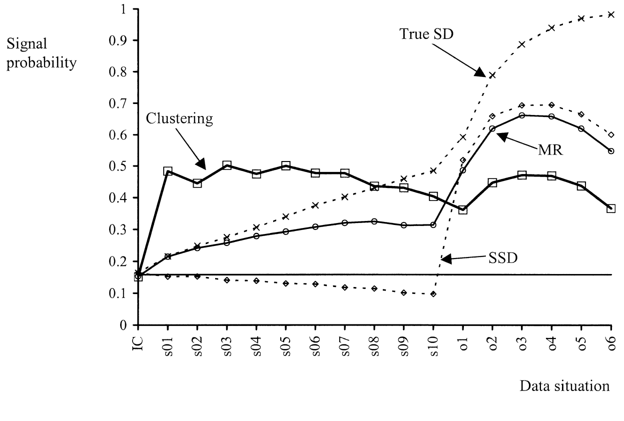

FIGURE 9. Signal Probability for X-Chart and Clustering Algorithm.

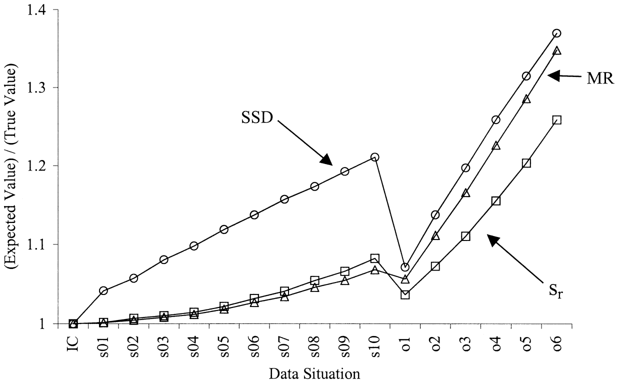

FIGURE 10. Comparison of Standard Deviation Estimators Under Various OC Situations.

dard deviation and the average of the moving ranges divided by d_2 in Figure 10. The sample standard deviation is the one most inflated by any of the OC situations. The performance of the average of the moving ranges is nearly as bad as the sample standard deviation in the presence of outliers. The clustering estimator s_r is fairly robust to outliers and nearly as good as the moving range estimator with the presence of shifts.

Conclusion

An algorithm has been described for detecting the presence of, and identifying the likely location(s) of, one or more shifts in the mean and/or outliers in individual observations. The main advantage of the proposed decision rule is the more uniform detection probability with respect to different out-of-control situations. While a CUSUM chart has greater detection probability for a single shift, it is less likely to detect multiple shifts and unlikely to detect the presence of any number of outliers. On the other hand, an X-chart does not do as well in detecting shifts, although it does better with outliers. In retrospective analysis, where multiple shifts and/or outliers may be present, the clustering algorithm gives a computationally simple way to detect the presence of shifts and/or outliers in any reasonable quantity. Furthermore, the clustering algorithm can be generalized to effectively detect shifts in the variance, trends, or for a variety of other OC conditions.

Acknowledgments

The author gratefully acknowledges the helpful suggestions of the Editor, Dr. Douglas M. Hawkins of the University of Minnesota, Dr. Zachary G. Stoumbos of Rutgers University, and the anonymous reviewers.

References

Akaike, H. (1974). “A New Look at the Statistical Identification Model”. *IEEE Transactions on Automatic Control* 19, pp. 716–723.

Barry, D. and Hartigan, J. A. (1993). “A Bayesian Analysis for Change Point Problems”. *Journal of the American Statistical Association* 88, pp. 309–319.

Chernoff, H. and Zacks, S. (1964). “Estimating the Current Mean of a Normal Distribution Which is Subjected to Changes in Time”. *Annals of Mathematical Statistics* 35, pp. 999–1028.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. (1977). “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. *Journal of the Royal Statistical Society* B 39, pp. 1–38.

Diebold, F. X. (2001). *Elements of Forecasting,* 2nd ed. South-Western, Cincinnati, OH.

George, E. I. (2001). “The Variable Selection Problem”. *Journal of the American Statistical Association* 95, pp. 1304–1308.

Giannini, C. (1992). *Topics in Structural VAR Econometrics*. Springer-Verlag, New York, NY.

Hawkins, D. M. (1976). “Point Estimation of the Parameters of Piecewise Regression Models”. *Applied Statistics* 25, pp. 51–57.

Hawkins, D. M. (2001). “Fitting Multiple Change-Point Models to Data”. *Computational Statistics and Data Analysis* 37, pp. 323–341.

Krishnaiah, P. R. and Miao, B. Q. (1988). “Review About Estimation of Change Points” in *Handbook of Statistics, Vol. 7,* edited by P. R. Krishnaiah and C. R. Rao, Elsevier Science, pp. 375–402.

Lai, T. L. (1995). “Change Point Detection in Quality Control and Dynamical Systems”. *Journal of the Royal Statistical Society* B 57, pp. 613–658.

Litterman, R. (1986) “Forecasting with Bayesian Vector Autoregressions—Five Years of Experience”. *Journal of Business and Economic Statistics* 4, pp. 25–38.

McGee, V. E. and Carleton, W. T. (1970) “Piecewise Regression”. *Journal of the American Statistical Association* 65, pp. 1109–1124

Quandt, R. E. (1958). “The Estimation of the Parameter of a Linear Regression System Obeying Two Separate Regimes”. *Journal of the American Statistical Association* 53, pp. 873–880.

Quandt, R. E. (1960). “Tests of the Hypotheses that a Linear Regression System Obeys Two Separate Regimes”. *Journal of the American Statistical Association* 55, pp. 324–330.

Quandt, R. E. (1972). “A New Approach to Estimating Switching Regressions”. *Journal of the American Statistical Association* 67, pp. 306–310.

Schwarz, G. (1978). “Estimating the Dimension of a Model”. *Annals of Statistics* 6, pp. 261–264.

Sullivan, J. H. (2002). “Estimating the Locations of Multiple Change Points in the Mean”. *Computational Statistics,* to appear.

SULLIVAN, J. H. and WOODALL, W. H. (1998). "Adapting Control Charts for the Preliminary Analysis of Multivariate Observations". *Communications in Statistics—Simulation and Computation* 27, pp. 953–979.

SULLIVAN, J. H. and WOODALL, W. H. (2000). "Change-Point Detection of Mean Vector or Covariance Matrix Shifts Using Multivariate Individual Observations". *IIE Transactions* 32, pp. 537–549.

VOSTRIKOVA, L. J. (1981). "Detecting 'disorder' in Multidimensional Random Processes". *Soviet Mathematics Doklady* 24, pp. 55–59.

YAO, Y. (1988). "Estimating the Number of Change-Points Via Schwarz Criterion". *Statistics and Probability Letters* 6, pp. 181–189.

ZACKS, S. (1991). "Detection and Change-Point Problems" in *Handbook of Analysis* edited by B. K. Ghosh and P. K. Sen, Marcel Dekker, New York, NY, pp. 531–562.

Key Words: *Quality control, Robust Estimation, Statistical Process Control.*